Attorney Docket No. ~~18547-018570~~ 056297-5012-01

PATENT APPLICATION

MYCOBACTERIAL *RpoB* SEQUENCES

Inventor: Thomas Gingeras
Jorg Drenkow

Assignee: AFFYMETRIX, INC.
3380 Central Expressway
Santa Clara, California 95051
a Corporation of California

Entity: Large

~~TOWNSEND and TOWNSEND and CREW LLP~~
~~Two Embarcadero Center, 8th Floor~~
~~San Francisco, California 94111-3834~~
~~(415) 326-2400~~

PATENT

Attorney Docket No. ~~18547-018570~~ 056297-5012-01

MYCOBACTERIAL *RpoB* SEQUENCES

STATEMENT OF GOVERNMENT INTEREST

[0001] The work described in this application was supported in part by grant number 1R43a140400 by the NIAID. The Government may have certain rights in this invention.

CROSS-REFERENCE TO RELATED APPLICATIONS

[0002] This application derives priority from USSN 60/080,616, filed April 3, ~~1999~~ 1998, and incorporated by reference. Applications USSN 08/797,812, filed February 7, 1997, now US Patent 6,228,575; USSN 60/011,339, filed ~~08~~ Feb. February 8, 1996; USSN 60/012,631, filed [~~01~~] March 1 1996; USSN 08/629,031, filed ~~08~~ April 8, 1996, now abandoned; and 60/017,765, filed 15 May 15, 1996 are directed to related subject matter. These applications are specifically incorporated by reference in their entirety for all purposes.

BACKGROUND OF THE INVENTION

Field of the Invention

[0003] This invention is directed to polymorphisms in rpoB genes of mycobacteria and use of the same in the identification and characterization of microorganisms.

Background of the Invention

[0004] Multidrug resistance and human immunodeficiency virus (HIV-1) infections are factors which have had a profound impact on the tuberculosis problem. An increase in the frequency of *Mycobacterium tuberculosis* strains resistant to one or more anti-mycobacterial agents has been reported, Block, et al., (1994) JAMA **271**:665-671. Immunocompromised HIV-1 infected patients not infected with *M. tuberculosis* are frequently infected with *M. avium* complex (MAC) or *M. avium-M. intracellulare* (MAI) complex. These mycobacteria species are often resistant to

the drugs used to treat *M. tuberculosis*. These factors have re-emphasized the importance for the accurate determination of drug sensitivities and mycobacteria species identification.

[0005] In HIV-1 infected patients, the correct diagnosis of the mycobacterial disease is essential since treatment of *M. tuberculosis* infections differs from that called for by other mycobacteria infections, Hoffner, S.E. (1994) *Eur. J. Clin. Microbiol. Inf. Dis.* **13**:937-941. Non-tuberculosis mycobacteria commonly associated with HIV-1 infections include *M. kansasii*, *M. xenopi*, *M. fortuitum*, *M. avium* and ~~*M. intracellular*~~ *M. intracellulare*, Wolinsky, E., (1992) *Clin. Infect. Dis.* **15**:1-12, Shafer, R.W. and Sierra, M.F. 1992 *Clin. Infect. Dis.* **15**:161-162. Additionally, 13% of new cases (HIV-1 infected and non-infected) of *M. tuberculosis* are resistant to one of the primary anti-tuberculosis drugs (isoniazid [INH], rifampin [RIF], streptomycin [STR], ethambutol [EMB] and pyrazinamide [PZA] and 3.2% are resistant to both RIF and INH, Block, et al., *JAMA* **271**:665-671, (1994). Consequently, mycobacterial species identification and the determination of drug resistance have become central concerns during the diagnosis of mycobacterial diseases.

[0006] Methods used to detect, and to identify *Mycobacterium* species vary considerably. For detection of *Mycobacterium tuberculosis*, microscopic examination of acid-fast stained smears and cultures are still the methods of choice in most microbiological clinical laboratories. However, culture of clinical samples is hampered by the slow growth of mycobacteria. A mean time of four weeks is required before sufficient growth is obtained to enable detection and possible identification. Recently, two more rapid methods for culture have been developed involving a radiometric, Stager, C.E. et al., (1991) *J. Clin. Microbiol.* **29**:154-157, and a biphasic (broth/agar) system Sewell, et al., (1993) *J. Clin. Microbiol.* **29**:2689-2472. Once grown, cultured mycobacteria can be analyzed by lipid composition, the use of species specific antibodies, species specific DNA or RNA probes and PCR-based sequence analysis of 16S rRNA gene (Schirm, et al. (1995) *J. Clin. Microbiol.* **33**:3221-3224; Kox, et al. (1995) *J. Clin. Microbiol.* **33**:3225-3233) and IS6110 specific repetitive sequence analysis (For a review see, e.g., Small et al., P.M. and van Embden, J.D.A. (1994) *Am. Society for Microbiology*, pp. 569-582). The analysis of 16S rRNA sequences (RNA and DNA) has been the most informative molecular approach to identify *Mycobacteria* species (Jonas, et al., *J. Clin. Microbiol.* **31**:2410-

2416 (1993)). However, to obtain drug sensitivity information for the same isolate, additional protocols (culture) or alternative gene analysis is necessary.

[0007] To determine drug sensitivity information, culture methods are still the protocols of choice. *Mycobacteria* are judged to be resistant to particular drugs by use of either the standard proportional plate method or minimal inhibitory concentration (MIC) method. However, given the inherent lengthy times required by culture methods, approaches to determine drug sensitivity based on molecular genetics have been recently developed.

[0008] Because resistance to RIF in *E. coli* strains was observed to arise as a result of mutations in the *rpo*B gene, Telenti, et al., id., identified a 69 base pair (bp) region of the *M. tuberculosis rpo*B gene as the locus where RIF resistant mutations were focused. Kapur, et al., (1995) *Arch. Pathol. Lab. Med.* **119**:131-138, identified additional novel mutations in the *M. tuberculosis rpo*B gene which extended this core region to 81 bp. In a detailed review on antimicrobial agent resistance in mycobacteria, Musser (*Clin. Microbiol. Rev.*, **8**:496-514 (1995)), summarized all the characterized mutations and their relative frequency of occurrence in this 81 bp region of *rpo*B. Missense mutations comprise 88% of all known mutations while insertions (3 or 6 bp) and deletions (3, 6 and 9 bp) account for 4% and 8% of the remaining mutations, respectively. Approximately 90% of all RIF resistant tuberculosis isolates have been shown to have mutations in this 81 bp region. The remaining 10% are thought possibly to involve genes other than *rpo*B.

[0009] For the above reasons, it would be desirable to have simpler methods which identify and characterize microorganisms, such as *Mycobacteria,* both at the phenotypic and genotypic level. This invention fulfills that and related needs.

SUMMARY OF THE INVENTION

[0010] In one aspect, the invention provides isolated nucleic acids comprising at least 25, 50, 75, 100, or 200 contiguous bases from an rpoB sequence shown in Table 1 (SEQ ID NOS: 1-181). Some nucleic acid comprise a complete sequence shown in Table 1.

[0011] The invention further provides a set of probes perfectly complementary to and spanning such nucleic acids, preferably spanning one of the complete sequences shown in Table 1 (SEQ ID NOS: 1-181).

[0012] The invention further provides methods of classifying mycobacteria. Some such methods entail providing a sample comprising a mycobacterial rpoB target nucleic acid from a mycobacteria, determining the sequence of a segment of at least 50 contiguous bases from the target nucleic acid; comparing the determined sequence to at least one sequence shown in Table 1; and classifying the mycobacteria from the extent of similarity of the compared sequences. Preferably, at least 100 or 200 contiguous bases are determined from the target nucleic acid. Preferably, the determined sequence is compared with a plurality of sequences from Table 1, for example, 10, 20, 50 or all of the sequence from Table 1 (SEQ ID NOS: 1-181).

[0013] In other methods of classification, the identity of one or more bases in the target sequence at one or more positions corresponding to one or more of the highlighted positions in a sequence shown in Table 1 is determined. The identity of the one or more bases characterizing the species of mycobacteria that is present in the sample. In some methods, the identity of at least 10 bases in the target nucleic acid at positions corresponding to highlighted positions in a sequence shown in Table 1 is determined. In some methods, the identity of at least 20 bases in the target sequence at highlighted positions shown in Table 1 are identified. In some methods, at least 20 determined bases are compared with 20 bases occupying corresponding positions in each of at least ten sequences from Table 1.

[0014] In another aspect, the invention provides sequence-specific polynucleotide probes or primers that hybridizes to a segment of a mycobacterial rpoB sequence shown in Table 1 or its complement without hybridizing to the M. tuberculosis sequence designated ATCC9-Mtb in Table 1 or its complement, the segment including a highlighted nucleotide position shown in Table 1. In some such probes, a central position of the probe aligns with a highlighted nucleotide position shown in ~~Tables 1~~ Table 1. In some such primers, the 3' end of the primer aligns with a highlighted nucleotide position shown in Table 1. Some probes and primers are between 10 and 50 bases long.

[0015] In another aspect, the invention provides a computer-readable storage medium for storing data for access by an application program being executed on a data processing system. Such a system comprises a data structure stored in the computer-readable storage medium. The data structure includes information resident in a database used by the application program and includes a plurality of records, each record comprising information identifying a polymorphism

or sequence shown in Table 1. Some records have a field identifying a base occupying a polymorphic site and a field identifying location of the polymorphic site. Some records record a contiguous segment of at least 50, 100, or 200 bases from an rpoB sequence shown in Table 1. Some storage medium comprise at least ten records each recording a contiguous segment of at least 50 bases from at least ten rpoB sequences shown in Table 1.

BRIEF DESCRIPTION OF THE DRAWINGS

[0016] Fig. 1: Computer that may be utilized to execute software embodiments of the present invention.

[0017] Fig. 2: A system block diagram of a typical computer system that may be used to execute software embodiments of the invention.

DEFINITIONS

[0018] A polynucleotide can be DNA or RNA, and single- or double-stranded. Polynucleotide can be naturally occurring or synthetic, and can be of any length. Preferred polynucleotide probes of the invention include contiguous segments of DNA, or their complements including any of the highlighted bases shown in Table 1. The segments are usually between 5 and 100 bases, and often between 5-10, 5-20, 10-20, 10-50, 20-50 or 20-100 bases. The highlighted site can occur within any position of the segment. Preferred polynucleotide probes are capable of binding in a base-specific manner to a complementary strand of nucleic acid. Such probes include peptide nucleic acids, as described in Nielsen et al., *Science* 254, 1497-1500 (1991), and probes having nonnaturally occurring bases.

[0019] The term primer refers to a single-stranded polynucleotide capable of acting as a point of initiation of template-directed DNA synthesis under appropriate conditions (*i.e.*, in the presence of four different nucleoside triphosphates and an agent for polymerization, such as, DNA or RNA polymerase or reverse transcriptase) in an appropriate buffer and at a suitable temperature. The appropriate length of a primer depends on the intended use of the primer but typically ranges from 15 to 30 nucleotides. Short primer molecules generally require cooler temperatures to form sufficiently stable hybrid complexes with the template. The term primer site refers to the area of the target DNA to which a primer hybridizes. The term primer pair

means a set of primers including a 5' upstream primer that hybridizes with the 5' end of the DNA sequence to be amplified and a 3', downstream primer that hybridizes with the complement of the 3' end of the sequence to be amplified.

[0020] A cDNA or cRNA is derived from an RNA if it produced by a process in which the RNA serves as a template for production of the cDNA or cRNA.

[0021] Hybridizations are usually performed under stringent conditions, for example, at a salt concentration of no more than 1 M and a temperature of at least 25°C. For example, conditions of 5 x SSPE (750 mM NaCl, 50 mM Na Phosphate, 5 mM EDTA, pH 7.4) and a temperature of 25-30°C are suitable for allele-specific probe hybridizations.

[0022] An isolated nucleic acid means an object species invention that is the predominant species present (*i.e.*, on a molar basis it is more abundant than any other individual species in the composition). Preferably, an isolated nucleic acid comprises at least about 50, 80 or 90 percent (on a molar basis) of all macromolecular species present. Most preferably, the object species is purified to essential homogeneity (contaminant species cannot be detected in the composition by conventional detection methods).

[0023] For sequence comparison and homology determination, typically one sequence acts as a reference sequence to which test sequences are compared. When using a sequence comparison algorithm, test and reference sequences are input into a computer, subsequence coordinates are designated, if necessary, and sequence algorithm program parameters are designated. The sequence comparison algorithm then calculates the percent sequence identity for the test sequence(s) relative to the reference sequence, based on the designated program parameters.

[0024] Optimal alignment of sequences for comparison can be conducted, *e.g.*, by the local homology algorithm of Smith & Waterman, *Adv. Appl. Math.* 2:482 (1981), by the homology alignment algorithm of Needleman & Wunsch, *J. Mol. Biol.* 48:443 (1970), by the search for similarity method of Pearson & Lipman, *Proc. Nat'l. Acad. Sci. USA* 85:2444 (1988), by computerized implementations of these algorithms (GAP, BESTFIT, FASTA, and TFASTA in the Wisconsin Genetics Software Package, Genetics Computer Group, 575 Science Dr., Madison, WI), or by visual inspection (*see generally*, Ausubel *et al., infra*).

[0025] One example of algorithm that is suitable for determining percent sequence identity and sequence similarity is the BLAST algorithm, which is described in Altschul *et al., J. Mol.*

Biol. 215:403-410 (1990). Software for performing BLAST analyses is publicly available through the National Center for Biotechnology Information ~~(http://www.ncbi.nlm.nih.gov/)~~. This algorithm involves first identifying high scoring sequence pairs (HSPs) by identifying short words of length W in the query sequence, which either match or satisfy some positive-valued threshold score T when aligned with a word of the same length in a database sequence. T is referred to as the neighborhood word score threshold (Altschul *et al.*, *supra*). These initial neighborhood word hits act as seeds for initiating searches to find longer HSPs containing them. The word hits are then extended in both directions along each sequence for as far as the cumulative alignment score can be increased. Cumulative scores are calculated using, for nucleotide sequences, the parameters M (reward score for a pair of matching residues; always > 0) and N (penalty score for mismatching residues; always < 0). For amino acid sequences, a scoring matrix is used to calculate the cumulative score. Extension of the word hits in each direction are halted when: the cumulative alignment score falls off by the quantity X from its maximum achieved value; the cumulative score goes to zero or below, due to the accumulation of one or more negative-scoring residue alignments; or the end of either sequence is reached. The BLAST algorithm parameters W, T, and X determine the sensitivity and speed of the alignment. The BLASTN program (for nucleotide sequences) uses as defaults a wordlength (W) of 11, an expectation (E) of 10, a cutoff of 100, M=5, N=-4, and a comparison of both strands. For amino acid sequences, the BLASTP program uses as defaults a wordlength (W) of 3, an expectation (E) of 10, and the BLOSUM62 scoring matrix (*see* Henikoff & Henikoff (1989) *Proc. Natl. Acad. Sci. USA* 89:10915).

[0026] In addition to calculating percent sequence identity, the BLAST algorithm also performs a statistical analysis of the similarity between two sequences (*see, e.g.*, Karlin & Altschul (1993) *Proc. Nat'l. Acad. Sci. USA* 90:5873-5787). One measure of similarity provided by the BLAST algorithm is the smallest sum probability (P(N)), which provides an indication of the probability by which a match between two nucleotide or amino acid sequences would occur by chance. For example, a nucleic acid is considered similar to a reference sequence if the smallest sum probability in a comparison of the test nucleic acid to the reference nucleic acid is less than about 0.1, more preferably less than about 0.01, and most preferably less than about 0.001.

[0027] The term "target nucleic acid" refers to a nucleic acid (often derived from a biological sample), to which the probe nucleic acid is designed to specifically hybridize. It is the presence or expression level of the target nucleic acid that is to be detected or quantified. The target nucleic acid has a sequence that is complementary to the nucleic acid sequence of the corresponding probe directed to the target. The term target nucleic acid may refer to the specific subsequence of a larger nucleic acid to which the probe is directed or to the overall sequence (*e.g.* gene or mRNA) whose expression level it is desired to detect. The difference in usage will be apparent from context.

[0028] "Subsequence" refers to a sequence of nucleic acids that comprise a part of a longer sequence of nucleic acids.

DETAILED DESCRIPTION

I. Mycobacterial Sequences of rpoB Genes

[0029] Table 1 shows a comparison of a substantial collection of mycobacterial strains of an about 700-nucleotide conserved region of an rpoB gene. The sequences shown in Table 1 are identified as follows: SEQ ID NOS: 1-56, respectively, are shown on pages 21, 25, 29, 33, 37, 41, 45, 49, 53, 57, 61 and 65; SEQ ID NOS: 57-112, respectively, are shown on pages 22, 26, 30, 34, 38, 42, 46, 50, 54, 58, 62 and 65; SEQ ID NOS: 113-168, respectively, are shown on pages 23, 27, 31, 35, 39, 43, 47, 51, 55, 59, 63 and 66; SEQ ID NOS: 169-181, respectively, are shown on pages 24, 28, 32, 36, 40, 44, 52, 56, 60, 64 and 68. The first sequence, designated as a reference sequence, is from *M. tuberculosis*. Nucleotides are numbered consecutively starting from the first nucleotide of the reference sequences. Other sequences are from other strains of mycobacteria. For example, the sequences designated ATCC-av, M29, M30...M104 are from *M. avium*. Sequences designated from ATT-chelnew, M11, M13, and M17 are from *M. chelonae*. Sequences designated ATCC-for, M53, M55, M56, and M74 are from *M. fortuitum*, and so ~~fourth~~ forth. Complete correspondence between strain designations and strain types is shown in Table 2. Nucleotides in a mycobacterial sequence are accorded the same number as the corresponding position of the reference sequence when the two are maximally aligned. Differences between a sequence and the reference sequences are shown in highlighted type. Many of the highlighted positions are common to all tested members of a species. Other

highlighted positions vary among different isolates in a species. Both types of variation can be useful in speciation analysis.

II. Analysis of Species Variations

A. Preparation of Samples

[0030] An rpoB sequence is isolated from a sample of an unknown mycobacteria being tested. Nucleic acids can be isolated from myobacteria by standard methods as described in WO 97/29212 (incorporated by reference in its entirety for all purposes). The rpoB sequences to be analyzed can then be isolated and amplified by means of PCR. *See generally PCR Technology: Principles and Applications for DNA Amplification* (ed. H.A. Erlich, Freeman Press, NY, NY, 1992); *PCR Protocols: A Guide to Methods and Applications* (eds. Innis, et al., Academic Press, San Diego, CA, 1990); Mattila et al., *Nucleic Acids Res.* 19, 4967 (1991); Eckert et al., *PCR Methods and Applications* 1, 17 (1991); *PCR* (eds. McPherson et al., IRL Press, Oxford); and U.S. Patent 4,683,202 (each of which is incorporated by reference for all purposes). Primers for PCR preferably flank the regions of interest rpoB genes, although primers to internal sites can be used if it is intended to analyze only certain sites of potential species variation. Exemplary primers are described in WO 97/29212. If necessary, additional sequences flanking the sequences shown in Table 1 can be determined using probes based on the sequences in Table 1 to isolate full-length rpoB sequences from the appropriate mycobacterial species.

B. Detection of Species-Specific Variations in Target DNA

1. Sequence-Specific Probes

[0031] The design and use of sequence-specific probes for analyzing polymorphisms is described by e.g., Saiki et al., *Nature* 324, 163-166 (1986); Dattagupta, EP 235,726, Saiki, WO 89/11548. Sequence-specific probes can be designed that hybridize to a segment of target DNA in one isolate of mycobacteria that do not isolate to a corresponding isolate in another due to the presence of allelic or species variations in the respective segments from the two sequences. Hybridization conditions should be sufficiently stringent that there is a significant difference in hybridization intensity between alleles, and preferably an essentially binary response, whereby a probe hybridizes to only one of the sequences. Some probes are designed to hybridize to a

segment of target DNA such that the site of potential sequence variation aligns with a central position (e.g., in a 15 mer at the 7 position; in a 16 mer, at either the 8 or 9 position) of the probe. This design of probe achieves good discrimination in hybridization between different allelic and species variants.

[0032] Sequence-specific probes are often used in pairs, one member of a pair showing a perfect match to a reference form of a target sequence and the other member showing a perfect match to a variant form. Several pairs of probes can then be immobilized on the same support for simultaneous analysis of multiple potential variations within the same target sequence.

2. Tiling Arrays

[0033] The bases occupying sites of potential variation can also be identified by hybridization to nucleic acid arrays, some example of which are described by WO 95/11995 (incorporated by reference in its entirety for all purposes). Such arrays contain a series of overlapping probes spanning a reference sequence. Any of the rpoB sequences shown in Table 1, or contiguous segments of, for example, at least 25, 50, 100 or 200 bases thereof, can serve as a reference sequence. WO 95/11995 also describes subarrays that are optimized for detection of a variant forms of a precharacterized polymorphism. Such a subarray contains probes designed to be complementary to a second reference sequence, which is a variant of the first reference sequence. The inclusion of a second group (or further groups) can be particular useful for analyzing short subsequences of the primary reference sequence in which multiple mutations are expected to occur within a short distance commensurate with the length of the probes (*i.e.*, two or more mutations within 9 to 21 bases).

3. Sequence-Specific Primers

[0034] A sequence-specific primer hybridizes to a site on target DNA overlapping a polymorphism and only primes amplification of a variant form to which the primer exhibits perfect complementarily. See Gibbs, *Nucleic Acid Res.* 17, 2427-2448 (1989). This primer is used in conjunction with a second primer which hybridizes at a distal site. Amplification proceeds from the two primers leading to a detectable product signifying the particular variant form is present. A control is usually performed with a second pair of primers, one of which

shows a single base mismatch at the site of variation and the other of which exhibits perfect complementarily to a distal site. The single-base mismatch prevents amplification and no detectable product is formed. The method works best when the mismatch is included in the 3'-most position of the primer aligned ~~with the~~ with the point of variation because this position is most destabilizing to elongation from the primer. *See, e.g.,* WO 93/22456.

4. Direct-Sequencing

[0035] The direct analysis of mycobacterial sequences can be accomplished using either the dideoxy chain termination method or the Maxam Gilbert method (see Sambrook et al., *Molecular Cloning, A Laboratory Manual* (2nd Ed., CSHP, New York 1989); Zyskind et al., *Recombinant DNA Laboratory Manual*, (Acad. Press, 1988)).

III. Methods of Use

[0036] The sequences and polymorphisms shown in Table 1 are useful for identifying the presence of myobacteria in samples, and optionally, classifying the mycobacteria. The sample can be obtained from a patient or from a biological source, such as a food product.

[0037] The sequences shown in Table 1 can be used for design of sequence-specific probes or primers encompassing polymorphic sites as described above. These probes or primers can then be used to determine the base occupying a corresponding position in an rpoB sequence from an isolate in a sample under test. A base in one sequence corresponds with a base in another when the two bases occupy the same position when the two sequences are maximally aligned by one of the criteria described in Definitions.

[0038] Alternatively, the sequences shown in Table 1 can be used for design of tiling arrays in which one or more of the sequences serves as a reference sequence. At least one set of overlapping probes is designed spanning a segment of the reference sequence, as described in WO95/11995 or EP 717,113. Target sequences from samples under test can be hybridized to such arrays, optionally in combination with controls of known rpoB sequences. The hybridization pattern of a target sequence to such an array can be analyzed to determine the identity of bases at which the target sequence differs from the reference sequence, as described in WO 95/11995.

[0039] One or more of the above methods, or direct sequencing, can be used to identify the base occupying at least one and usually several (e.g., 5, 10, 15, 25, 50 or 100) sites of potential variation between the 16S RNA and/or rpoB gene in an unknown mycobacteria relative to bases occupying corresponding sites in one or more known strains of mycobacteria, such as those shown in ~~Tables 1~~ Table 1. This analysis results in a profile of bases occupying particular sites that characterizes the mycobacterial strain under test. The profile is compared with the corresponding profiles of different mycobacterial isolates shown in e.g., ~~Tables 1~~ Table 1. In general, the unknown mycobacterium isolate is characterized as being from the same mycobacterial species as the precharacterized isolate with which it shares the greatest similarity in base profile.

[0040] In some methods, the sequence of a contiguous segment of the rpoB target nucleic acid is determined in a sample under test for comparison with one or more of the sequences shown in Table 1. The mycobacteria is classified by the extent of similarity. For example, if a target nucleic acid shows greater sequence identity to rpoB sequences from one species than any other, the sample from which the target was obtained is typically classified as arising from that species.

[0041] Alternatively, an array of tiled probes based on a reference sequence shown in Table 1 can be used for identifying and characterizing mycobacterial sequences based on comparison of hybridization patterns. Such an array is hybridized to a 16S RNA or rpoB target sequence from a sample, and the hybridization pattern compared with the hybridization pattern of one or more control sequences. The hybridization patterns of control sequences can be historic controls, stored, for example, in a computer database, or can be contemporaneous controls performed at or near the same time as the hybridization to the target sequence. Optionally, hybridization of target and reference sequence can be performed simultaneously using different labels.

[0042] Method of classifying unknown mycobacterial isolate by matching the hybridization pattern of a target sequence with those of control sequences from characterized species are described in more detail in WO 97/29212 (incorporated by reference in its entirety for all purposes). In an idealized case, the detection of a particular hybridization pattern in an isolate characterizes that isolate as belonging to a particular species. This can occur when the hybridization pattern detected in the isolate is uniquely associated with a specific species. More frequently however, such an unique one-to-one correspondence is not present. Instead, the

hybridization pattern observed in an isolate does not bear a unique correspondence with a previously characterized species. However, the hybridization pattern detected is associated with a probability of the organism being screened belonging to a particular species (or not) or carrying a particular phenotypic trait (or not). As a result, analysis of an increasing number of polymorphic sites in an isolate, allows one to classify the isolated with an increasing level of confidence. Algorithms can be used to derive such composite probabilities from the comparison of multiple polymorphic forms between an isolate and references. Typically, the mathematical algorithm makes a call of the identity of the species and assign a confidence level to that call. One can determine the confidence level (>90%, >95% etc.) that one desires and the algorithm will analyze the hybridization pattern and either provide an identification or not. Occasionally, the call is that the sample may be one of two, three or more species, in which case a specific identification is not be possible. However, one of the strengths of this technique is that the rapid screening made possible by the chip-based hybridization allows one to continuously expand a database of patterns ultimately to enable the identification of species previously unidentifiable due to lack of sufficient information.

IV. Modified Polypeptides and Gene Sequences

[0043] The invention further provides variant forms of nucleic acids and corresponding proteins. The nucleic acids comprise one of the sequences described in ~~Tables 1~~ Table 1. Some nucleic acid encode full-length variant forms of proteins. Variant proteins have the prototypical amino acid sequences of encoded by nucleic acid sequence shown in ~~Tables 1~~ Table 1 (read so as to be in-frame with the full-length coding sequence of which it is a component).

[0044] Variant genes can be expressed in an expression vector in which a variant gene is operably linked to a native or other promoter. Usually, the promoter is a eukaryotic promoter for expression in a mammalian cell. The transcription regulation sequences typically include a heterologous promoter and optionally an enhancer which is recognized by the host. The selection of an appropriate promoter, for example trp, lac, phage promoters, glycolytic enzyme promoters and tRNA promoters, depends on the host selected. Commercially available expression vectors can be used. Vectors can include host-recognized replication systems,

amplifiable genes, selectable markers, host sequences useful for insertion into the host genome, and the like.

[0045] The means of introducing the expression construct into a host cell varies depending upon the particular construction and the target host. Suitable means include fusion, conjugation, transfection, transduction, electroporation or injection, as described in Sambrook, *supra*. A wide variety of host cells can be employed for expression of the variant gene, both prokaryotic and eukaryotic. Suitable host cells include bacteria such as *E. coli*, yeast, filamentous fungi, insect cells, mammalian cells, typically immortalized, *e.g.*, mouse, CHO, human and monkey cell lines and derivatives thereof. Preferred host cells are able to process the variant gene product to produce an appropriate mature polypeptide. Processing includes glycosylation, ubiquitination, disulfide bond formation, general post-translational modification, and the like.

[0046] The protein may be isolated by conventional means of protein biochemistry and purification to obtain a substantially pure product, *i.e.*, 80, 95 or 99% free of cell component contaminants, as described in Jacoby, *Methods in Enzymology* Volume 104, Academic Press, New York (1984); Scopes, *Protein Purification, Principles and Practice*, 2nd Edition, Springer-Verlag, New York (1987); and Deutscher (ed), *Guide to Protein Purification, Methods in Enzymology,* Vol. 182 (1990). If the protein is secreted, it can be isolated from the supernatant in which the host cell is grown. If not secreted, the protein can be isolated from a lysate of the host cells.

[0047] In addition to substantially full-length polypeptides expressed by variant genes, the present invention includes biologically active fragments of the polypeptides, or analogs thereof, including organic molecules which simulate the interactions of the peptides. Biologically active fragments include any portion of the full-length polypeptide which confers a biological function on the variant gene product, including ligand binding, and antibody binding. Ligand binding includes binding by nucleic acids, proteins or polypeptides, small biologically active molecules, or large cellular structures.

[0048] Polyclonal and/or monoclonal antibodies that specifically bind to variant gene products but not to corresponding prototypical gene products are also provided. Antibodies can be made by injecting mice or other animals with the variant gene product or synthetic peptide fragments thereof. Monoclonal antibodies are screened as are described, for example, in Harlow & Lane,

Antibodies, A Laboratory Manual, Cold Spring Harbor Press, New York (1988); Goding, *Monoclonal antibodies, Principles and Practice* (2d ed.) Academic Press, New York (1986). Monoclonal antibodies are tested for specific immunoreactivity with a variant gene product and lack of immunoreactivity to the corresponding prototypical gene product. These antibodies are useful in diagnostic assays for detection of the variant form, or as an active ingredient in a pharmaceutical composition.

V. Kits

[0049] The invention further provides kits comprising at least one sequence-specific probe as described above. Often, the kits contain one or more pairs of sequence-specific probes hybridizing to different forms of a polymorphism. In some kits, the sequence-specific probes are provided immobilized to a substrate. For example, the same substrate can comprise sequence-specific probes for detecting at least 10, 100 or all of the variations shown in Table 1. Optional additional components of the kit include, for example, restriction enzymes, reverse-transcriptase or polymerase, the substrate nucleoside triphosphates, means used to label (for example, an avidin-enzyme conjugate and enzyme substrate and chromogen if the label is biotin), and the appropriate buffers for reverse transcription, PCR, or hybridization reactions. Usually, the kit also contains instructions for carrying out the methods.

VI. Computer Databases

[0050] Fig. 1 illustrates an example of a computer system that can be used to store records relating to polymorphisms of the invention and perform algorithms comparing polymorphic profiles and to classify species. ~~Fig. 32~~ Fig. 2 shows a computer system 100 which includes a monitor 102, screen 104, cabinet 106, keyboard 108, and mouse 110. Mouse 110 may have one or more buttons such as mouse buttons 112. Cabinet 106 houses a CD-ROM drive 114, a system memory and a hard drive (see ~~Fig. 33~~ Fig. 2) which can be utilized to store and retrieve software programs incorporating code that implements the present invention, data for use with the present invention, and the like. Although a CD-ROM 116 is shown as an exemplary computer readable storage medium, other computer readable storage media including floppy disks, tape, flash

memory, system memory, and hard drives may be utilized. Cabinet 106 also houses familiar computer components such as a central processor, system memory, hard disk, and the like.

[0051] Fig. 2 shows a system block diagram of computer system 100 that may be used to execute software embodiments of the present invention. As in ~~Fig. 32~~ Fig. 1, computer system 100 includes monitor 102 and keyboard 108. Computer system 100 further includes subsystems such as a central processor 102, system memory 120, I/O controller 122, display adapter 124, removable disk 126 (e.g., CD-ROM drive), fixed disk 128 (e.g., hard drive), network interface 130, and speaker 132. Other computer systems suitable for use with the present invention may include additional or fewer subsystems. For example, another computer system can include more than one processor 102 (i.e., a multi-processor system) or a cache memory.

[0052] Arrows such as 134 represent the system bus architecture of computer system 100. However, these arrows are illustrative of any interconnection scheme serving to link the subsystems. For example, a local bus can be utilized to connect the central processor to the system memory and display adapter. Computer system 100 shown in ~~Fig. 33~~ Fig. 1 is but an example of a computer system suitable for use with the present invention.

[0053] The computer stores records relating to the polymorphisms of the record. Some such records record a polymorphism by reference to the position of a polymorphic site and the identity of base(s) occupying that site in one or more species. Some databases include records for at least ten polymorphic sites in at least ten of the sequences shown in ~~Tables 1~~ Table 1. Some databases include records for all of the polymorphic sites in at least one of the sequences shown in ~~Tables 1~~ Table 1. Some databases includes records for at least 100, 1000, or 2000 polymorphic sites shown in ~~Tables 1~~ Table 1. Some databases include records for all of the polymorphic sites shown in ~~Tables 1~~ Table 1.

[0054] The foregoing invention has been described in some detail by way of illustration and example, for purposes of clarity and understanding. It will be obvious to one of skill in the art that changes and modifications may be practiced within the scope of the appended claims. Therefore, it is to be understood that the above description is intended to be illustrative and not restrictive. The scope of the invention should, therefore, be determined not with reference to the above description, but should instead be determined with reference to the following appended claims, along with the full scope of equivalents to which such claims are entitled.

[0055] All patents, patent applications and publications cited in this application are hereby incorporated by reference in their entirety for all purposes to the same extent as if each individual patent, patent application or publication were so individually denoted.

MYCOBACTERIAL *Rpo*B SEQUENCES

ABSTRACT OF THE DISCLOSURE

[0056] This invention provides polynucleotide probes, sequences and methods for speciating and phenotyping organisms, for example, using probes based on the *Mycobacterium tuberculosis rpo*B gene. The groups or species to which an organism belongs may be determined by comparing hybridization patterns of target nucleic acid from the organism to hybridization patterns in a database.

Attorney Docket No. 056297-5012-01

PATENT APPLICATION

MYCOBACTERIAL *RpoB* SEQUENCES

Inventor: Thomas Gingeras

Jorg Drenkow

Assignee: AFFYMETRIX, INC.

3380 Central Expressway

Santa Clara, California 95051

a Corporation of California

Entity: Large